Pseudorandom sets and explicit constructions of Ramsey graphs

Pavel Pudlák*
Mathematical Institute of the Academy of Sciences, and
Institute of Theoretical Computer Science†
Prague, Czech Republic

Vojtěch Rödl‡
Emory University
Atlanta, Georgia,USA

February 16, 2004

Abstract

We shall show a polynomial time construction of a graph G on N vertices such that neither G nor $\overline{G}$ contains $K_{r,r}$, for $r = \sqrt{N}/2^{\sqrt{\log N}} = o(\sqrt{N})$. To this end we construct a subset $X \subseteq \mathbb{F}_2^m$ which has small intersections with all subspaces of dimension $m/2$.

1 Introduction

In 1947 Erdös proved that there are graphs on N vertices which do not contain a clique or independent set of size $(2 + o(1)) \log_2 N$. This was one of the first applications of the probabilistic method in combinatorics, a method by which one can prove the existence of a finite structure without finding a concrete definition of it. Therefore he asked if such graphs, possibly with a constant larger than 2, can be defined explicitly [4, 3]. This problem is still open; the best result so far is due to Frankl and Wilson [6]. Frankl and Wilson found an explicit construction of graphs that do not contain a clique or independent set of size $2^{\Theta(\sqrt{\log N \log\log N})}$. In this paper we shall consider a related problem about bipartite graphs. Let $G = (U, V, E)$, $E \subseteq U \times V$ be a bipartite graph, let $|U| = |V| = N$. For such graphs it is well-known that there exist two sets $A \subseteq U$ and $B \subseteq V$ such that either $A \times B \subseteq E$ or $A \times B \cap E = \emptyset$ and $|A| = |B| = k = (1 - o(1)) \log_2 N$. In other words either G or its complement $\overline{G}$ contains a copy of $K_{k,k}$. A probabilistic argument again gives the existence

*Partially supported by grant A1019401 of the AV ČR.
†Project No. LN00A056 of the Ministry of Education of the Czech Republic.
‡Partially supported by grant DMS0300529 of the NSF.

of $N \times N$ bipartite graphs G such that neither G nor its complement $\overline{G}$ contains a copy of $K_{k,k}$ for $k = (2 + o(1)) \log_2 N$. The explicit construction problem for bipartite graphs seems to be more difficult than for ordinary symmetric graphs. This is apparent from the fact that there are no known constructions of bipartite graphs for which one can prove that neither they nor their complement contains $K_{k,k}$ for $k = o(\sqrt{N})$. For $k = \sqrt{N}$ such graphs can be constructed from Hadamard matrices; the proof uses Lindsey's Lemma (see, e.g., [5]). Notice that the problem is not only to find a "good candidate" for such a graph, but also to prove that it has the property. It has been conjectured that the Paley graphs (for prime fields) provide an example of such graphs for some $k = O(\log N)$, but the present proof techniques allow us only to prove it for $k = \Theta(\sqrt{N})$.

In this paper we shall be concerned with the problem of constructing such graphs. First we shall give a reduction of this problem to the construction of randomly looking subspaces of $\mathbb{F}_2^n$ ($\mathbb{F}_2$ denotes the two-element field). Then we shall present our construction. It is not explicit in the way that would satisfy Erdös. We shall show that the graph can be constructed in polynomial time in the size of the graph. Our construction is based on derandomizing an existence proof of suitable pseudorandom sets on a small support and expanding them to large ones in order to obtain a polynomial time bound. Thus it does not have a simple compact description.

The problem of constructing such subspaces is interesting *per se*, because it can be viewed as a problem of constructing generators that are pseudorandom with respect to linear tests. Therefore, in Section 4 we shall survey some related results and show connections to other problems. A simple observation enables us to improve some results on hitting sets and lower bounds on bounded depth circuits.

We shall conclude the paper with suggesting number-theoretical construction that might achieve better parameters and which are quite explicit. For these construction we can only prove the square root bound, but we feel that for suitable fields one should get more. Also it seems to us that they are more manageable than the Paley construction, since it is well-known that the problem of the distribution of quadratic residues is very hard. What is needed for our constructions is to estimate the sizes of intersections of very simple curves over $\mathbb{F}_{2^n}$ with linear spaces over $\mathbb{F}_2$.

2 The basic construction

Our starting point is the the inner product graph $H_n = (\mathbb{F}_2^n, E_n)$ where

$$(u, v) \in E_n \leftrightarrow u \cdot v = 1,$$

(we use dot for the inner product modulo 2). Let $N = 2^n$ denote the number of vertices of H_n. It is well-known that neither H_n nor $\overline{H}_n$ contains K_k for $k = (1 + o(1))\sqrt{N}$. In [10] we showed that this graphs contains an induced subgraph G on $\sqrt{N}$ vertices which is Ramsey, meaning that neither G nor its complement contains K_k for $k = (2 + o(1)) \log_2 N$. The same can be said about the bipartite version of H_n (this is the bipartite graph (U_n, V_n, F_n) in which the vertices of U_n and V_n are indexed by the vectors of $\mathbb{F}_2^n$, and two vertices indexed by u and v are connected iff $u \cdot v = 1$). Again the bipartite version of H_n contains an induced subgraph G such that neither G nor its (bipartite) complement $\overline{G}$ contains $K_{k,k}$ for $k = (2 + o(1)) \log_2 N$. This suggests the problem to find such a subgraph explicitly. As it is an induced subgraph we are only looking for a suitable subset S of $\mathbb{F}_2^n$.

For $S \subseteq \mathbb{F}_2^n$, we let G_S be the graph obtained from the bipartite version of H_n by restricting it to the vertices indexed by elements of S. (Thus G_S is the bipartite graph (U_S, V_S, F_S) where $U_S = \{u_i \; ; \; i \in S\}$, $V_S = \{v_i \; ; \; i \in S\}$, and $(u_i, v_j) \in F_S$ iff $i \cdot j = 1$.) What we need is stated in the following proposition.

Proposition 1 *Suppose every vector space $V \subseteq \mathbb{F}_2^n$ of dimension $\lfloor (n+1)/2 \rfloor$ intersects S in less than r elements, then neither G_S nor the bipartite complement $\overline{G}_S$ contains $K_{r,r}$.*

Proof. We shall argue by contradiction. Let X and Y be subsets of S of size r, and suppose they determine an induced subgraph of G_S isomorphic to $K_{r,r}$ or $\overline{K}_{r,r}$. Thus for some $\alpha \in \{0, 1\}$, $x \cdot y = \alpha$ for every $x \in X$ and $y \in Y$. Interpreting vectors y as solutions of the system of linear equations determined by the vectors x and the constant α, we get $\dim\langle X \rangle + \dim\langle Y \rangle^a \leq n$, where $\langle X \rangle$ denotes the vector space spanned by X and $\langle Y \rangle^a$ denotes the affine span of Y. Since every affine subspace is contained in a vector subspace whose dimension is larger by at most 1, we have $\dim\langle X \rangle + \dim\langle Y \rangle \leq n + 1$. Whence either $\dim\langle X \rangle \leq (n+1)/2$ or $\dim\langle Y \rangle \leq (n+1)/2$. Thus, by the assumption of the lemma, either $|\langle X \rangle \cap S| < r$ or $|\langle Y \rangle \cap S| < r$. But $X, Y \subseteq S$ and $|X| = |Y| = r$, hence this is a contradiction. ∎

Thus we have reduced the Ramsey problem to the problem of finding an explicit set $S \subseteq \mathbb{F}_2^n$ that has small intersections with all $\lfloor (n+1)/2 \rfloor$-dimensional vector spaces. As we observed in the proof of the lemma, it does not matter if we consider affine subspaces instead of vector subspaces. Let us stress, however, that the main problem is the explicitness of the set S, because a random set of size $2^{n/2}$ has intersections of size $(2 + o(1)) \log N$. So we want to find an explicit construction of sets that share some properties of random sets, namely, they 'look random for linear sets'.

3 A construction

In this section we shall show a polynomial time construction of a set X such that the intersections with spaces of dimension $m/2$ are asymptotically smaller than $\sqrt{|X|}$. By Proposition 1 this gives a polynomial time construction of a bipartite graph G_X with sets of vertices of size $N = 2^m$ such that neither G_X nor $\overline{G}_X$ contains $K_{r,r}$ for $r = o(\sqrt{N})$.

More precisely we prove the following theorem.

Theorem 2 *Let k be an odd natural number, $n = 66k$ and $m = kn$. Then it is possible to construct a set $X \subseteq \mathbb{F}_2^m$, $|X| = 2^{m/2}$ in time $2^{O(m)}$ (hence polynomial time in the size of X) such that for every affine subspace $L \subseteq \mathbb{F}_2^n$ of dimension $m/2$*

$$|L \cap X| \leq \frac{\sqrt{|X|}}{2^{\sqrt{\log |X|}}}.$$

Corollary 3 *For infinitely many numbers N it is possible to construct a graph G on N vertices in time polynomial in N such that neither G nor $\overline{G}$ contains $K_{r,r}$ for $r = \sqrt{N}/2^{\sqrt{\log N}}$.*

Here and in the sequel log denotes the logarithm to the base 2 and ln is the natural logarithm.

The idea of the proof is to construct a $Y \subseteq \mathbb{F}_2^n$ for some small n by derandomizing the random choice of such a set by the method of conditional expectations. If n is small enough, then the running time will still be $2^{O(m)}$. Then we take the product of an odd number of copies of this set. The fact that an odd number cannot be evenly split will give us the improvement over the easy square root bound.

We shall start by proving the existence of a suitable set by a counting argument. We need to control intersections for all dimensions $d = 0, \dots, n$, because the copies of Y will be in a much larger vector space. We shall bound the logarithm of the size of the intersections by the following function:

$$h(x) =_{df} \max(2 \log x + 4, x - n/2 + 2),$$

for $1 \leq x \leq n$. Observe that the breaking point of the graph of the function h extended to all real numbers in the interval $[1, n]$, i.e., the number b such that $2 \log b + 4 = b - n/2 + 2$, satisfies $b = n/2 + 2 \log n + 1 + o(1)$. Thus $b < n/2 + 2 \log n + 2$, for n sufficiently large.

Lemma 4 *For all sufficiently large even numbers n, there exists $Y \subseteq \mathbb{F}_2^n$, $|Y| = 2^{n/2}$ such that for all vector subspaces V, if $V \cap Y \neq \emptyset$, then*

$$\log |V \cap Y| \leq h(\dim V).$$

Hence for all affine subspaces L, if $L \cap Y \neq \emptyset$, then

$$\log |L \cap Y| \leq h(\dim L) + 1.$$

Proof. We shall show that if n is sufficiently large, then for all $d \geq 1$ and all vector spaces V, $\dim V = d$, we have the following estimate for a random Y of size $2^{n/2}$:

$$\Pr(\log |V \cap Y| > h(d)) \cdot 2^{nd} < 1/n. \tag{1}$$

Since 2^{nd} is an upper bound on the number of vector subspaces of dimension d, it follows that there exists a set Y required in the lemma. Notice that for d above the breaking point b and $0 \leq t \leq n - d$, we have $h(d + t) = h(d) + t$. Also every space of dimension $d + t$ is a disjoint union of 2^t spaces of dimension d. Hence if for such a d the bound $\log |V \cap Y| \leq h(d)$ holds for all spaces of dimension d, it also holds for all $d' \geq d$. Thus it suffices to prove (1) for d specified in the following three cases.

Case 1 $n/2 + 2 \log n \leq d < n/2 + 2 \log n + 2$

For a fixed $V \subseteq \mathbb{F}_2^n$ with $\dim V = d$ and a randomly chosen subset $Y \subseteq \mathbb{F}_2^n$ with $|Y| = 2^{n/2}$ the random variable $|V \cap Y|$ has hypergeometric distribution with parameters 2^n, 2^d, and

$2^{n/2}$. Hence, the expectation $\mathbf{E}(|V \cap Y|) = \lambda = 2^{d-n/2}$. Consequently, by the inequality (see, e.g., [8, Theorem 2.8])

$$\Pr(|V \cap Y| \geq \lambda + t) \leq \exp\left(-\frac{t^2}{2\lambda} + \frac{t^3}{6\lambda^2}\right) \tag{2}$$

we infer

$$\begin{aligned} \Pr(\log |V \cap Y| > h(d)) \leq \Pr(\log |V \cap Y| > d - n/2 + 2) \leq \\ \Pr(|V \cap Y| \geq 3 \times 2^{d-n/2}) \leq \exp\left(-\frac{2}{3} 2^{d-n/2}\right). \end{aligned} \tag{3}$$

Thus for n sufficiently large,

$$\begin{aligned} 2^{nd} \exp\left(-\frac{2}{3} 2^{d-n/2}\right) \leq 2^{n(n/2+2\log n+2)} \exp\left(-\frac{2}{3} 2^{2\log n}\right) = \\ \exp(n^2(\ln 2/2 - 2/3) + 2\ln 2(n \log n + 1)) < \frac{1}{n}, \end{aligned}$$

which yields (1) for this case.

Case 2 $n/2 \leq d \leq n/2 + 2\log n$

Recall that $h(x) = \max(h_1(x), h_2(x))$, where $h_1(x) = 2\log x + 4$ and $h_2(x) = x - n/2 + 2$. Notice that in Case 1 we have shown (1) for h_2 instead of h (see (3)). Also we have $h(n/2) = h_2(n/2 + 2\log n)$, whence $h(d) = h_1(d) \geq h_1(n/2) \geq \lceil h_2(n/2 + 2\log n) \rceil$ for $n/2 \leq d \leq n/2 + 2\log n$. Thus if we let V be a d-dimensional subspace for an arbitrary d in the interval $[n/2, n/2 + 2\log n]$ and let V' be a $\lceil n/2 + 2\log n \rceil$-dimensional space containing V, then we have

$$\Pr(\log |V \cap Y| > h(d)) \leq \Pr(\log |V' \cap Y| > h_2(\lceil n/2 + 2\log n \rceil)) \leq \frac{1}{n 2^{n(\lceil n/2 + 2\log n \rceil)}} \leq \frac{1}{n 2^{nd}}.$$

Case 3 $1 \leq d < n/2$

Again $h(d) = 2\log d + 4$. In this case we apply the inequality

$$\Pr(|V \cap Y| \geq c 2^{d-n/2}) \leq \left(\frac{c}{e}\right)^{-c 2^{d-n/2}},$$

(c.f. in [8]) with $c = 2^{2\log d + 4 - d + n/2}$ to infer that

$$\Pr(|V \cap Y| \geq 2^{h(d)}) \leq \left(\frac{c}{e}\right)^{-16d^2}.$$

Since

$$2^{nd} \left(\frac{c}{e}\right)^{-16d^2} \leq 2^{(-n/d + n/2 - d + 2\log d)16d^2} < \frac{1}{n}$$

for every d considered above we infer that (1) also holds in this case.

■

Lemma 5 *For $1 \leq \ell \leq k \leq N$, the binomial coefficient $\binom{k}{\ell}$ can be computed in time $N^{O(1)}$.*

Proof. Using the Pascal triangle the problem is reduced to $O(N^2)$ additions of natural numbers of length at most N. ■

Corollary 6 *Let $0 \leq k, s \leq 2^{n/2}$, $1 \leq d \leq n$, $a_1, \ldots, a_k \in \mathbb{F}_2^n$. Let $Y = \{y_1, \ldots, y_{2^{n/2}}\}$ be a random subset of $\mathbb{F}_2^n$ of size $2^{n/2}$. Then a rational number*

$$q = \sum_{V,\ \dim V = d} \Pr(|V \cap Y| > s \mid y_1 = a_1, \ldots, y_k = a_k)$$

can be computed in time $2^{O(n)}$.

Proof. Given $a_1, \ldots, a_k \in \mathbb{F}_2^n$ with $|\{a_1, \ldots, a_k\} \cap V| = \ell$, we have

$$q = \frac{\sum_{t>s} \binom{2^d - \ell}{t-l} \binom{2^n - 2^d - k + \ell}{2^{n/2} - t - k + \ell}}{\binom{2^n - k}{2^{n/2} - k}}.$$

Applying Lemma 5 with $N = 2^n$ and adding over at most 2^n summands yields our claim. ■

Lemma 7 *A set Y satisfying the conditions of Lemma 4 can be constructed using a deterministic algorithm in time $2^{O(n^2)}$.*

Proof. We shall apply the method of conditional expectations. For $k = 0, 1, \ldots, 2^{n/2} - 1$, let

$$P_k(V, s, a_1, \ldots, a_k) =_{df} \Pr(|V \cap Y| > s \mid y_1 = a_1, \ldots, y_k = a_k).$$

In the proof of Lemma 4 we showed

$$\sum_{d=1}^{n} \sum_{\dim V = d} P_0(V, 2^{h(d)}) < 1.$$

Since

$$P_k(V, s, a_1, \ldots, a_k) = \mathbf{E}_x P_{k+1}(V, s, a_1, \ldots, a_k, x),$$

we have

$$\sum_{d=1}^{n} \sum_{\dim V = d} P_k(V, 2^{h(d)}, a_1, \ldots, a_k) = \mathbf{E}_x \sum_{d=1}^{n} \sum_{\dim V = d} P_{k+1}(V, 2^{h(d)}, a_1, \ldots, a_k, x).$$

Hence, given $a_1, \ldots, a_k$, we can find a_{k+1} such that

$$\sum_{d=1}^{n} \sum_{\dim V = d} P_{k+1}(V, 2^{h(d)}, a_1, \ldots, a_k, a_{k+1}) \leq \sum_{d=1}^{n} \sum_{\dim V = d} P_k(V, 2^{h(d)}, a_1, \ldots, a_k)$$

by searching through at most 2^n elements of $\mathbb{F}_2^n \setminus \{a_1, \ldots, a_k\}$. Since the sum has $\sum_{d=1}^{n} 2^{nd} = 2^{O(n^2)}$ summands, this requires time $\leq 2^n \cdot 2^{O(n^2)} = 2^{O(n^2)}$. Hence in time $\leq 2^{n/2} \cdot 2^{O(n^2)} = 2^{O(n^2)}$ we find $a_1, \ldots, a_{n/2}$ such that

$$\sum_{d=1}^{n} \sum_{\dim V = d} P_{n/2}(V, 2^{h(d)}, a_1, \ldots, a_{2^{n/2}}) < 1.$$

Since $P_{2^{n/2}}$ is the characteristic function of the relation $|V \cap Y| > s$, we are done. ■

Proof. (of Theorem 2)

Recall that $n = 66k$, $m = nk$ and set $W = \mathbb{F}_2^n$. Take $Y \subseteq W$ from Lemma 7. It can be constructed in time $2^{O(n^2)} = 2^{O(m)}$. We shall identify $\mathbb{F}_2^m$ with W^k. So we can define $X = Y^k \subseteq \mathbb{F}_2^m$. This is the construction.

Now we shall estimate the size of $V \cap X$ for a subspace $V \subseteq \mathbb{F}_2^m$. We are interested only in subspaces of dimension $m/2$, but the following argument is general. Let $d = \dim V$. We can represent V as the tree

$$T_V =_{df} \{(y_1, \ldots, y_i) \ ; \ i = 0, \ldots, k, \ y_1, \ldots, y_i \in W, \ \exists y_{i+1}, \ldots, y_k \in W \ (y_1, \ldots, y_k) \in V\}.$$

Given $(y_1, \ldots, y_i) \in T_V$, $i < k$, the set of the successors of $(y_1, \ldots, y_i)$ in T_V

$$L_{y_1 \ldots y_i} =_{df} \{y_{i+1} \ ; \ (y_1, \ldots, y_{i+1}) \in T_V\}$$

is an affine subspace of W. To see this, let $e_i : V \to W$ denote the i-th coordinate map and let $\pi_i : V \to W^i$ be the projection onto the first i coordinates. For $(y_1, \ldots, y_i) \in T_V$, choose $v = (y_1, \ldots, y_i, \ldots, y_k) \in V$. Then

$$L_{y_1 \ldots y_i} = e_{i+1}\left(\pi_i^{-1}(y_1, \ldots, y_i)\right) = e_{i+1}(v + \ker \pi_i) = e_{i+1}(v) + e_{i+1}(\ker \pi_i)$$

is an affine space of W of dimension $\dim e_{i+1}(\ker \pi_i)$, which is the same for all $(y_1, \ldots, y_i) \in T_V$. Call this dimension d_{i+1}. Writing v_i for $\pi_i(V)$, we have $\dim V_{i+1} = \dim V_i + d_{i+1}$, and hence $\sum_{i=0}^{k} d_i = \dim V = d$.

Since $(y_1, \ldots, y_k) \in X$ iff every $y_i \in Y$, we can estimate

$$|V \cap X| \leq \prod_{i=0}^{k-1} \max_{(y_1, \ldots, y_i) \in T_V} |L_{y_1 \ldots y_i} \cap Y|.$$

Due to the property of Y, $\log |L_{y_1 \ldots y_i} \cap Y| \leq h(d_i) + 1$, and hence

$$|V \cap X| \leq \prod_{i=1}^{k} 2^{h(d_i)+1} = 2^{\sum_{i=1}^{k} (h(d_i)+1)}.$$

Hence it remains to estimate $\max \sum_{i=1}^{k} h(d_i)$ for $d_1, \ldots, d_k$ such that $\sum_{i=1}^{k} d_i = m/2$, which is the content of the following lemma.

Lemma 8 *For $0 \leq d_1, \ldots, d_k \leq n$ such that $\sum_{i=1}^k d_i = m/2$,*

$$\max \sum_{i=1}^{k} h(d_i) \leq m/4 - n/8 + k - 1.$$

Before proving the lemma, we shall finish the proof of the theorem. Since $m = nk$ and $n = 66k$, we have $k = \sqrt{m/66}$ and $n = \sqrt{66m}$. Thus

$$\max \sum_{i=1}^{k} (h(d_i) + 1) \leq m/4 - n/8 + 2k - 1 = m/4 - \sqrt{66m}/8 + 2\sqrt{m}/\sqrt{66} - 1 =$$

$$m/4 - \frac{25}{4\sqrt{66}}\sqrt{m} - 1 = \frac{1}{2}\log|X| - \frac{25}{4\sqrt{33}}\sqrt{\log|X|} - 1 =$$

$$\frac{1}{2}\log|X| - 1.087\ldots\sqrt{\log|X|} - 1 < \frac{1}{2}\log|X| - \sqrt{\log|X|},$$

for m sufficiently large. ■

Proof. (of Lemma 8)

We extend the definition of the function h to all real numbers $0 \leq x \leq n$ by defining $h(x) = 4x$ for $0 \leq x \leq 1$ and using the original formula for $1 \leq x \leq n$. We shall prove the lemma for all real $0 \leq d_1, \ldots, d_k \leq n$ and for the extended h.

Let $0 \leq d_1, \ldots, d_k \leq n$ be such that $\sum_{i=1}^k d_i = m/2$ and $\sum_{i=1}^k h(d_i)$ is maximal. Recall that b denotes the breaking point of h near $n/2$ and that $b = x_0 = n/2 + \log n + o(1)$ with the quantity $o(1)$ positive; so $n/2 + \log n < b < n/2 + \log n + 1$. Consider the possible positions of d_i in the intervals $[0, b]$ and $[b, n]$.

1. W.l.o.g. we can assume that all $d_i \in (b, n]$ are equal to n, except possibly for one. Indeed, if there are two $d_i, d_j \in (b, n)$, $d_i \leq d_j$, then we can gradually decrease d_i and increase d_j until one of them hits an end-point of $[d, n]$. Let ℓ be the number of the elements d_i that are equal to n. Since $kn = m$ and k is odd, $\ell \leq \frac{k-1}{2}$.

2. The number $\frac{2}{\ln 2}$ is the unique $x \in (0, b)$ such that $h'(x) = 1$; $h'(x) > 1$ for $x < \frac{2}{\ln 2}$, $x \neq 1$, and $h'(x) < 1$ for $\frac{2}{\ln 2} < x < b$.

Claim. *If some $d_i \in (b, n)$, then all $d_j \in [0, b]$ are equal to $\frac{2}{\ln 2}$.*

Proof. Suppose it were not true. Let $d_i \in (b, n)$ and, $d_j > \frac{2}{\ln 2}$. Then, for small $\epsilon > 0$, $d_i := d_i + \epsilon$ and $d_j := d_j - \epsilon$ would produce larger value of $\sum_{i=1}^k h(d_i)$. If $d_j < \frac{2}{\ln 2}$, then move the points in the opposite directions.

3. We shall use this claim to prove that in fact: *there is no $d_i \in (b, n)$.*

Proof. Suppose $d_i \in (b, n)$. Consider two cases. If $\ell = (k-1)/2$, then, by 2.,

$$m/2 = n(k-1)/2 + d_i + \frac{2}{\ln 2}(k-1)/2,$$

whence

$$d_i = m/2 - n(k-1)/2 - \frac{2}{\ln 2}(k-1)/2 = n/2 - \frac{2}{\ln 2}(k-1)/2 < b,$$

which is a contradiction. If $\ell \leq (k-1)/2 - 1$, then, by 2.,

$$m/2 = n\ell + d_i + (k-\ell-1)\frac{2}{\ln 2} \leq n((k-1)/2-1) + n + k\frac{2}{\ln 2} = n(k-1)/2 + k\frac{2}{\ln 2} < nk/2 = m/2,$$

because $k\frac{2}{\ln 2} < 33k = n/2$. Thus we have a contradiction again.

4. Since h is convex on $[0, b]$, then there exists $a \in [0, b]$ such that all $b_i \in [0, b]$ are equal to a. Thus we have ℓ elements equal to n and $k - l$ elements equal to a.

5. Observe that 64 is the unique number $x \in [0, b]$ such that $h(x) = x/4$; thus $h(x) \leq x/4$, for $x \in [64, b]$.

Claim. *If n is sufficiently large, then $a \geq 64$.*
Proof. Suppose that $a < 64$. Then we have

$$m/2 = n\ell + (k-\ell)a < n(k-1)/2 + 64(k+1)/2 =$$

$$kn/2 - n/2 + 64(k+1)/2 = m/2 - 66k/2 + 64(k+1)/2 < m/2,$$

which is a contradiction.

6. Finally, we can estimate $\max \sum_{i=1}^{k} h(d_i)$.

$$\max \sum_{i=1}^{k} h(d_i) \leq h(n)\ell + h(a)(k-\ell) \leq (\frac{n}{2} + 2)\ell + \frac{a}{4}(k-\ell) =$$

$$(\frac{n}{4} + 2)\ell + \frac{1}{4}(a(k-\ell) + n\ell) = (\frac{n}{4} + 2)\ell + \frac{1}{4}\sum_{i=1}^{k} d_i \leq$$

$$(\frac{n}{4} + 2)(k-1)/2 + m/8 = m/8 + \frac{n}{4} \cdot \frac{k}{2} - n/8 + k - 1 = m/4 - n/8 + k - 1.$$

■

4 Pseudorandom sets for linear tests.

In order to apply Proposition 1 we need a construction of a set S which has small intersection with vector spaces of a certain dimension. If we took a random set S instead, then it would have not only small intersections, but moreover we could show that the sizes of intersections are concentrated around a value that depends on the size (which is determined by the dimension) of the vector spaces. Thus we can consider a more difficult problem of constructing sets S which have intersections with vectors spaces of a given dimension of size close to the size of such intersections with a random set. We can generalize it further and take an arbitrary family $\mathcal{F}$ of subsets of a set A and ask the same question. This problem has been studied for various families $\mathcal{F}$. Several technical terms have been used, such as *approximation, discrepancy* and *pseudorandomness*. In this paper we shall use the last one.

In this section we shall show that some well-known construction achieve very good parameters of pseudorandomness not only with respect to combinatorial rectangles, but also for vector subspaces. Unfortunately, this concerns only a range of parameters that is not interesting for the problem of constructing Ramsey graphs. Thus the reader who is only interested in the main result stated in the abstract can safely skip this section.

Definition 1 *Let $\mathcal{F}$ be a set of subsets of a set A, let $S \subseteq A$. We say that*

- *S is ε-*pseudorandom *for $\mathcal{F}$, if for all $W \in \mathcal{F}$,*

$$|\Pr(x \in W \mid x \in S) - \Pr(x \in W)| \leq \varepsilon,$$

where the probability is taken with respect to the uniform distribution on A (thus $\Pr(\ldots \mid x \in S)$ is the probability with respect to the uniform distribution on S);

- *S is* hitting *for $\mathcal{F}$, if for all $W \in \mathcal{F}$,*

$$S \cap W \neq \emptyset;$$

- *S is r-*evasive *for $\mathcal{F}$, if for all $W \in \mathcal{F}$,*

$$|S \cap W| \leq r.$$

The concepts of pseudorandom and hitting sets are well-known, the concept of evasive sets is new.

Let us rewrite the inequality defining the ε-pseudorandomness of S as follows

$$\left| |S \cap W| - |S| \frac{|W|}{|A|} \right| \leq \varepsilon |S|.$$

The meaning of this is that the intersection $|S \cap W|$ differs from the expected intersection, which we would get if S were random, by at most $\varepsilon|S|$. Thus the concept of pseudorandom sets is the strongest one, as it can be used to prove both the hitting property and the evasiveness. However constructions of pseudorandom sets with good parameters ε are known only for very special families (see [11]).

In the rest of this paper we will consider only set systems $\mathcal{F}$ that are all affine subspaces of $\mathbb{F}_2^n$ of some fixed dimension d. Such a set system will be denoted by $\mathcal{L}_2^{d,n}$. Notice that *ε-pseudorandom* for $\mathcal{L}_2^{n-1,n}$ is also called *2ε-biased*, [1]. As our aim is to look for constructions of sets that would give better constructive bounds on Ramsey theorem, we shall call a set S *evasive* if it is $o(\sqrt{|S|})$-evasive for $\mathcal{L}_2^{n/2,n}$. Note that we do not specify the size of S, it can be large, and it can be small. However, we surely want S to be computable in polynomial time in its size, which may be more difficult to show if S is small.

For applying Proposition 1 it suffices to have sets that are sufficiently evasive only for *vector subspaces*, but, since every affine set of dimension d is contained in a vector space of dimension $d+1$, r-evasive set for vector spaces of dimension $d+1$ is also r-evasive for $\mathcal{L}_2^{d,n}$, there is essentially no difference between the two concepts and we can concentrate on affine sets. (Similarly, ε-pseudorandomness for vector spaces of dimension d and $d+1$, implies 3ε-pseudorandomness for $\mathcal{L}_2^{d,n}$.)

Pseudorandom sets. We shall show that pseudorandomness is, up to a factor < 2, preserved downwards.[1] This is an easy consequence of Vazirani's lemma.

Lemma 9 (Vazirani, see [1]) *Let S be ε-pseudorandom multiset for $\mathcal{L}_2^{n-1,n}$ (= 2ε-biased). Then for every $k \leq n$, $1 \leq i_1 < \ldots < i_k \leq n$ and $a_{i_1}, \ldots, a_{i_k} \in \{0,1\}$,*

$$\left|\Pr(x_{i_1} = a_{i_1}, \ldots, x_{i_k} = a_{i_k} \mid (x_1, \ldots, x_n) \in S) - 2^{-k}\right| \leq 2(1 - 2^{-k})\varepsilon. \tag{4}$$

Corollary 10 *Let $1 \leq k < d < n$. If S is ε-pseudorandom for $\mathcal{L}_2^{d,n}$, then it is $2(1-2^{-k-1})\varepsilon$-pseudorandom for $\mathcal{L}_2^{d-k,n}$.*

Proof. 1. First assume $d = n-1$. Observe that Vazirani's lemma says that S is $2(1-2^{-k})\varepsilon$-pseudorandom for the set system of all combinatorial rectangles of co-dimension k. These sets are affine sets of co-dimension k of a special form. But any affine set of co-dimension k can be represented in this form if we suitably change the basis and changing basis has no influence on the property of being ε-pseudorandom for $\mathcal{L}_2^{n-1,n}$. Thus we get the corollary for $d = n-1$.

2. To get the general case, observe that if S is ε-pseudorandom for $\mathcal{L}_2^{d,n}$, $d < n$, then for every $L \in \mathcal{L}_2^{d+1,n}$ the set $S \cap L$ is $(\frac{2|S|}{|S \cap L|}\varepsilon)$-pseudorandom for affine subspaces K of L of co-dimension 1. Hence to estimate the intersection of S with a affine set K of dimension $d-k$ take an arbitrary $d+1$-dimensional affine space L containing K and apply the above fact. ∎

Let us now apply the above result to one of the constructions of [1], the *Powering Construction*:

$$C_n^{2m} =_{df} \{(x \cdot y, x^2 \cdot y, \ldots, x^n \cdot y) \ ; \ x, y \in \mathbb{F}_{2^m}\},$$

where $x \cdot y$ denotes the scalar product of x and y with x, y interpreted as elements of $\mathbb{F}_2^m$. We have, treating C_n^{2m} as a multiset,

- $|C_n^{2m}| = 2^{2m}$;
- C_n^{2m} is $(n-1)2^{-m}$ biased (ie., $(n-1)2^{-m-1}$ pseudorandom for $\mathcal{L}_2^{n-1,n}$).

Corollary 11 *For $1 \leq d < n$, C_n^{2m} is $(1-2^{-n+d})(n-1)2^{-m}$-pseudorandom for $\mathcal{L}_2^{d,n}$.*

One application of pseudorandom sets is in proving lower bounds on the size of circuits computing boolean functions. Suppose we have a class $\mathcal{C}$ of circuits which for a given n compute the sets $\mathcal{F} \subseteq \mathcal{P}(\{0,1\}^n)$. To find a set that cannot be computed by $\mathcal{C}$, we only need to construct a set $S \subseteq \{0,1\}^n$ which is sufficiently pseudorandom for $\mathcal{F}$. Namely, we need an ε-pseudorandom set $|S|$ with $\varepsilon < 1 - |S|/2^n$. A less trivial application was used in [7]. A set S is called an *ε-discriminator* for W, if

$$|\Pr(x \in W \mid x \in S) - \Pr(x \in W | \ x \notin S)| \geq \varepsilon.$$

[1] It is not preserved upwards: every set S is $2^d/|S|$-pseudorandom for $\mathcal{L}_2^{d,n}$, simply because the size of the sets in $\mathcal{L}_2^{d,n}$ is at most 2^d.

Lemma 12 ([7]) *Suppose a function f can be computed by a circuit $T_t^m(C_1,\dots,C_m)$, where T_t^m is the threshold function of m boolean variables and threshold t, and $C_1,\dots,C_m$ are some circuits. Then the set $S=\{a\in\{0,1\}^n\ ;\ f(a)=1\}$ is a $1/m$-discriminator for one of the sets $\{a\in\{0,1\}^n\ ;\ C_i(a)=1\}$, $i=1,\dots,m$.*

Hence if S is not an ε-discriminator for any of the sets computed by circuits from a class $\mathcal{C}$, then every circuit of the form $T_t^m(C_1,\dots,C_m)$ computing S, with $C_1,\dots,C_m\in\mathcal{C}$, must have $m\geq 1/\varepsilon$. This was used to prove an exponential lower bound on depth 2 threshold circuits with bounded weights.

Pseudorandomness is closely connected with discriminators. By a straightforward computation we get

$$\Pr(x\in W\mid x\in S)-\Pr(x\in W|\ x\notin S)=\frac{1}{\Pr(x\notin S)}(\Pr(x\in W\mid x\in S)-\Pr(x\in W)).$$

Whence, if S is ε-pseudorandom for $\mathcal{F}$, and $|S|<2^{n-1}$, then S is not a 2ε-discriminator for every $W\in\mathcal{F}$. Thus we can apply the above lemma with the circuits C_i computing affine spaces to prove a lower bound on the size of such circuits computing the set C_n^{2m}. Affine spaces are defined by circuits that are conjunctions of parities. Hence we can prove a lower bound on the size of circuits that have the following three levels: a threshold function on the top, ANDs in the middle level, and parities on the bottom. Taking $m=n/2-1$ we obtain the following result.

Corollary 13 *Let C be a depth three circuit of the type described above and suppose n is even. If C computes C_n^{n-2}, then the size of C is at least $2^{n/2}/(n-2)$.*

This improves an earlier result of Jukna [9], in which the top gate was only allowed to be an OR gate.

Hitting sets. Andreev, Clementi and Rolim [2] found an explicit hitting set of size $2^{O(k)}$ for $\mathcal{L}_2^{n-k,n}$, for $k\geq n^{2/3+o(1)}$. This is optimal up to the constant in the exponent. Their set is the set of vectors that encode boolean functions of $\log_2 n$ variables of circuit complexity $\leq ck/\log k$ for a suitable constant c (assume n is a power of two). For larger dimensions they construct a hitting set only for a restricted subset of $\mathcal{L}_2^{d,n}$. Using Corollary 11 we can show that the sets C_n^{2m} are hitting sets also for affine sets of larger dimensions and give more precise estimates on their size.

Corollary 14 *There is an explicit construction of a hitting set for $\mathcal{L}_2^{d,n}$ whose size is $\leq 2^{2(n-d+\log_2 n)}$. Namely, it is the set $C_n^{2(n-d+\log_2 n)}$.*

Proof. It suffices to show that $C_n^{2(n-d+\log_2 n)}$ is ε-pseudorandom with $\varepsilon<2^d/2^n$, which follows from Corollary 11. ■

Evasive sets. Constructions of sets with small bias (=pseudorandom for $\mathcal{L}_2^{n-1,n}$) do give us bounds for smaller dimensional sets, but not sufficiently good for evasiveness. It is because the minimal bias that one can achieve for a set of size $o(2^n)$ is only of the order of $1/\sqrt{|S|}$, see [1].

Another possible approach is to first construct evasive sets in a smaller space and then compose it somehow to get such a set on $\mathbb{F}_2^n$. Such an approach works very well in case of symmetric graphs. If we first derandomize the probabilistic construction on a smaller set of vertices and then use the lexicographic product to get a larger graph, we get almost the same bound as in the Frankl-Wilson construction. In case of evasive sets the most naive attempt is to take the product of small evasive sets. This does not work if the number of terms in the product is even, but, as we have shown in the previous section, we do get some gain if the number of terms is odd. One would expect that it should not be hard to find a considerably better construction than the product, but all the other construction that we tried failed.

5 Number-theoretical constructions

We propose to study sets of the form

$$S = \{(x, y) \ ; \ f(x, y) = 0\},$$

where $f(x, y) = 0$ is an algebraic equation over $\mathbb{F}_{2^{n/2}}$ (in (x, y) we interpret x and y as elements of $\mathbb{F}_2^n$, while in $f(x, y) = 0$ as elements of $\mathbb{F}_{2^{n/2}}$, and we assume that n is even).

We shall show that for $y = x^3$ and $xy = 1$, the set is $O(1/\sqrt{|S|})$-evasive for vector spaces of dimension $n/2$. This follows from the following observation.

Lemma 15 *If $X \subseteq S$, then $|X + X| \geq \binom{|X|}{2}$.*

Proof. Let $(x_1, y_1), \ldots, (x_4, y_4) \in S$, $(x_1, y_1) \neq (x_2, y_2)$, $(x_3, y_3) \neq (x_4, y_4)$ and assume

$$(x_1, y_1) + (x_2, y_2) = (x_3, y_3) + (x_4, y_4).$$

It follows that $x_1 + x_2 = x_3 + x_4 \neq 0$. Then we get easily $x_1 x_2 = x_3 x_4$. Thus $\{x_1, x_2\} = \{x_3, x_4\}$. Hence $\{(x_1, y_1), (x_2, y_2)\} = \{(x_3, y_3), (x_4, y_4)\}$, since x's determine y's. ∎

Now let W be a vector space of dimension $n/2$. The span of $S \cap W$ is contained in W and has size $\Omega(|S \cap W|^2)$. Hence $|S \cap W| = O(2^{n/4}) = O(\sqrt{|S|})$.

The same can be proved for S defined by $y^2 + cy = x^3 + ax + b$ (the finite points of an elliptic curve).

Problem 1 *Are these sets evasive?*

We only know that they are not evasive if $\mathbb{F}_{2^{n/2}}$ contains a subfield of size $2^{n/4}$. This fact suggests that a positive answer to this problem requires a nonelementary argument.

Notice that the two constructions above are also special cases of the following general construction

$$S = \{(x, \phi(x)) \ ; \ x \in \mathbb{F}_2^{n/2}\},$$

determined by a function $\phi : \mathbb{F}_2^{n/2} \to \mathbb{F}_2^{n/2}$. We know that for a random ϕ this works perfectly, so the problem can be phrased: to find a function that looks random to linear tests.[2]

Acknowledgment

We are grateful to Noga Alon for a useful discussion concerning the problems studied this paper, and to Eduardo Tengan and Mathias Schacht for many remarks improving the presentation.

References

[1] N. Alon, O. Goldreich, J. Håstad and R. Peralta, Simple constructions of almost k-wise independent random variables. *Random Structures and Algorithms,* 3(3), 1992, pp. 289-303.

[2] A. Andreev, A.E.F. Clementi and J.D.P. Rolim, Towards efficient constructions of hitting sets that derandomize BPP. *Electronic Colloquium on Computational Complexity TR96-029* (1996).

[3] F. Chung, R. Graham, *Erdös on Graphs, His Legacy of Unsolved Problems.* A K Peters, 1999.

[4] P. Erdös, Some remarks on the theory of graphs. *Bull. AMS* 53, 1947, pp.292-294.

[5] P. Frankl, V. Rödl and R. M. Wilson, *J. Combin. Theory Ser. B 44,* (1988), no. 3, 317–328;

[6] P. Frankl and R. M. Wilson, Intersection theorems with geometric consequences. *Combinatorica 1,* (1981), 259-286.

[7] A. Hajnal, W. Maass, P. Pudlák, M. Szegedy, G. Turán, Threshold circuits of bounded depth. *J. of Computer and System Science 46,* (1993), pp. 129-154.

[8] S. Janson, T. Łuczak and A. Ruciński, *Random graphs.* Wiley-Interscience, New York, 2000.

[9] S. Jukna, On graph complexity. *Electronic Colloquium on Computational Complexity TR04-005* (2004).

[10] P. Pudlák, V. Rödl, P. Savický, Graph complexity. *Acta Informatica 25,* (1988), pp.515-535.

[11] A. Srinivasan, Low-discrepancy sets for high-dimensional rectangles: a survey. *Bulletin EATCS 70*, (2000), pp. 67-76.

[2] One of the factorization algorithms, the *rho* algorithm, needs a "random" function. The experience suggests that very simple polynomials, such as x^2+1, work very well. In our case we cannot use a quadratic polynomial, since it defines a linear function over $\mathbb{F}_2$; the simplest nonlinear polynomial is x^3.